b) if it is determined that the candidate search result is a near-duplicate of another candidate search result, then rejecting the candidate search result.

47. (NEW) A search filter for processing search results to remove near-duplicates, the search filter comprising:

a) a near-duplicate determination facility for determining, for each of a predetermined number of candidate search results, whether the candidate search result is a near-duplicate of another candidate search result, and wherein the near-duplicate determination facility includes a comparison facility for comparing a cluster identifier of the candidate search result with that of another candidate search result, and wherein if the cluster identifiers of the two candidate search results match, then it is concluded that the two candidate search results are near-duplicates; and

b) a filter for rejecting the candidate search result if it is determined that the candidate search result is a near-duplicate of another candidate search result.

48. (NEW) A machine-readable medium having stored thereon a plurality of records, each of the records comprising:

a) a first field for storing a document identifier; and

b) a plurality of lists, each of the plurality of lists containing elements of a document identified by the document identifier stored in the first field,

wherein a hash function is used to determine which of the plurality of lists each of the elements will be contained in.

49. (NEW) A method for determining whether two documents are near-duplicates, the method comprising:

a) for each of the two documents, generating at least two fingerprints; and

b) determining whether or not the two documents are near-duplicate documents by

1) determining whether or not any one of the fingerprints of a first of the two documents matches any one of the fingerprints of a second of the two documents, and

2) if it is determined that any one fingerprint of the first of the two documents does match any one fingerprint of the second of the two documents, then concluding that the two documents are near-duplicates.

50. (NEW) A machine-readable medium having stored thereon a plurality of records, each of the records comprising:

a) a first field for storing a document identifier; and

b) a plurality of lists, each of the plurality of lists containing elements of a document identified by the document identifier stored in the first field,

wherein at least some of the plurality of lists include different numbers of elements.

51. (NEW) The machine-readable medium of claim 51 wherein at least one of the plurality of lists include no elements.

52. (NEW) A machine-readable medium having stored thereon a plurality of records, each of the records comprising:
a) a first field for storing a document identifier; and
b) a plurality of lists, each of the plurality of lists containing elements of a document identified by the document identifier stored in the first field,
wherein contiguous elements in a document are not necessarily contiguous elements of a list.

53. (NEW) A machine-readable medium having stored thereon a plurality of records, each of the records comprising:
a) a first field for storing a document identifier; and
b) a plurality of lists, each of the plurality of lists containing elements of a document identified by the document identifier stored in the first field,
wherein for each of the records, the number of lists is the same.

54. (NEW) The machine-readable medium of claim 54 wherein a number of the plurality of lists is independent of document size.

REMARKS

Please enter the foregoing amendments before examining this application.

New claims 46-49 correspond to claims 15 (in independent form), 31 (in independent form), 32 and 39, respectively, from U.S. Patent Application Serial No. 09/768,947 ("the parent application"), as filed. Since each of claims 15, 31, 32 and 39 was rejected in the parent

application, the reasons that the corresponding claims are allowable are set forth below.

New claims 46 and 47

In the parent application, claims 15 and 31 (corresponding to new claims 46 and 47, respectively) were rejected under 35 U.S.C. § 102 as being anticipated by U.S. Patent No. 6,119,124 ("the Broder patent"). The applicants believe that this rejection was incorrect, and that new claims 46 and 47 are allowable, in view of the following.

The Broder patent does not teach determining whether the candidate search result is a near-duplicate of another candidate search result by (1) comparing a cluster identifier of the candidate search result with that of the other candidate search result, and (2) if the cluster identifiers of the two candidate search results match, then concluding that the two candidate search results are near-duplicates. In the parent application, the Examiner argued that (i) the Broder patent teaches comparing a cluster identifier of a candidate search result with that of another candidate search result (citing column 6, lines 54-63 of the Broder patent) and (ii) the Broder patent teaches concluding that the two candidate search results are near duplicates if the cluster identifiers of the two candidate search results match (citing column 5, lines 13-20). Both of these assertions are incorrect.

First, the Broder patent does not teach comparing a cluster identifier of a candidate search result with that of another candidate search result. Column 6, lines 54-63 of the Broder patent discusses using fingerprints (i.e., short unique identifications of some sort of document information, such as shingles). This passage is not

concerned with cluster identifiers, nor does it concern determining whether cluster identifiers of two documents match. Accordingly, new claims 46 and 47 are not anticipated by the Broder patent for at least this reason.

Second, the Broder patent does not teach concluding that the two candidate search results are near duplicates if the cluster identifiers of the two candidate search results match. Column 5, lines 13-20 of the Broder patent discusses using document "features" (i.e., reduced "sketches") to determine whether documents are highly resembling (High resemblance is defined as occurring when documents have many common shingles (See column 6, lines 9-13.).), and if so, grouping the documents into the same cluster. This is the reverse of what is being claimed. That is, using document features to cluster documents is not the same as, and therefore does not teach, using cluster identifiers to determine whether documents are near duplicates. Accordingly, new claims 46 and 47 are not anticipated by the Broder patent for at least this additional reason.

New claim 48

In the parent application, claim 32 (corresponding to new claim 48) was rejected under 35 U.S.C. § 102 as being anticipated by the Broder patent. The applicants believe that this rejection was incorrect, and that new claim 48 is allowable, in view of the following.

The Broder patent does not teach using a hash function to determine which of a plurality of lists will contain a document element. In the parent application, the Examiner argued that the Broder patent teaches using a hash function to determine which of the plurality of lists each of the

elements will be contained in, citing column 6, lines 5-7. Column 6, lines 5-7 of the Broder patent describes the k-shingling of a document. Shingles are overlapping sets of a fixed number of document elements (or "tokens"). As will be illustrated by the following example, shingling is different than hashing.

Consider, for example, a document containing the elements A B A C D E B F G F A. Under the Broder patent, a four-shingling operation would produce eight lists:

ABAC
BACD
ACDE
CDEB
DEBF
EBFG
BFGF
FGFA

Depending on the hash functions used, the claimed invention might produce four lists:

AAFFA
D
BCB
EG

As can be appreciated from the foregoing example, when using shingling as taught by the Broder patent, (1) each list contains the same number of elements, (2) the elements in each list are contiguous in the original document, and (3) the number of lists is proportional to the length of

the document. On the other hand, when hashing is used to populate the lists, (1) lists may contain different numbers of elements (some lists can even be empty), (2) the elements in a list are not typically contiguous in the original document, and (3) the number of lists is fixed, regardless of the size of the document. To reiterate, shingling does not teach hashing. Accordingly, claim 48 is not anticipated by the Broder patent for at least this reason.

New claim 49

In the parent application, claim 39 (corresponding to new claim 49) was rejected under 35 U.S.C. § 102 as being anticipated by the Broder patent. The applicants believe that this rejection was incorrect, and that new claim 49 is allowable, in view of the following.

The Broder patent does not teach concluding that two documents are near-duplicates if <u>any one</u> fingerprint of one of the documents matches <u>any one</u> fingerprint of the other document. In the parent application, the Examiner argued that the Broder patent teaches this feature, citing column 5, lines 17-19. However, this section deals with clustering documents. Moreover, this section states that documents that share more than a predetermined number (which does not teach one, and in fact suggests more than one) of "features" are highly resembling. Even assuming, arguendo, that features as used in the Broder patent could be construed as fingerprints as claimed, determining high resemblance between documents if <u>more than a predetermined number</u> of features match does not teach determining documents to be near duplicates if <u>any one</u> of their

respective fingerprints match. Thus, claims 49 is not anticipated for at least this reason.

Claims 50-54

New claims 50-54 are similar to claim 48 but recite properties of the lists rather than how elements are assigned to the list. These claims are supported, for example, by Figures 5, 12A and 12B.

Conclusion

In view of the foregoing remarks, the applicants respectfully submit that this application is in condition for allowance.

Respectfully submitted,

June 27, 2003

John C. Pokotylo, Attorney
Reg. No. 36,242
Customer No. 26479
(732) 335-1222

STRAUB & POKOTYLO
1 Bethany Road
Suite 83, Bldg. 6
Hazlet, NJ 07730